International Journal of Modern Physics A


# DATA REPROCESSING ON WORLDWIDE DISTRIBUTED SYSTEMS

DANIEL WICKE
*Fermilab, P.O. Box 500*
*Batavia, Illinois 60510, USA**

Received (Day Month Year)
Revised (Day Month Year)

The DØ experiment faces many challenges in terms of enabling access to large datasets for physicists on four continents. The strategy for solving these problems on worldwide distributed computing clusters is presented. Since the beginning of Run II of the Tevatron (March 2001) all Monte-Carlo simulations for the experiment have been produced at remote systems. For data analysis, a system of regional analysis centers (RACs) was established which supply the associated institutes with the data. This structure, which is similar to the tiered structure foreseen for the LHC was used in Fall 2003 to reprocess all DØ data with a much improved version of the reconstruction software. This makes DØ the first running experiment that has implemented and operated all important computing tasks of a high energy physics experiment on systems distributed worldwide.

*Keywords*: Grid; Distributed Computing; Data Challenge.

## 1. Introduction

The improved understanding of the DØ detector lead to significant improvements of the physics reconstruction software. Its procedures and parameters are no longer based on design specification, but on the actual detector layout and its performance.

To benefit from these new developments $100\,\mathrm{pb}^{-1}$ of data processed with previous software releases had to be reprocessed with this new reconstruction software. As the on-site resources weren't sufficient to perform the task, computing resources of seven institutes from six countries were included in the effort to perform the first HEP data reprocessing on distributed, non-dedicated systems[1,2,3,4].

## 2. The Computing Task

The DØ reconstruction code reconstructs physics objects from raw data, i.e. the recorded digitised output of the detector. Reconstruction from raw data needs access to the calorimeter calibration constants which reside in a calibration database. It

*on leave of absence from Bergische Universität Wuppertal

Table 1. Computing centres participating. The computer power estimated to be available to DØ is given for each site in numbers of CPUs equivalent to a 1GHz Pentium-III processor.

| Centre | | Estimated CPUs |
|---|---|---|
| GridKa (Karlsruhe, Germany) | | 300 |
| IN2P3 (Lyon, France) | | 220 |
| Nikhef (Amsterdam, Netherlands) | local | 80 |
| | EDG | 320 |
| SAR (Texas, USA) | | 130 |
| UK (London, Manchester, Rutherford) | | 340 |
| WestGrid (Canada) | | 300 |

produces two main result files: The DST keeps most of the raw information for detailed studies, while the TMB is a condensed format aimed at physics analyses. In addition to the actual results metadata describing each of these files are created. These contain information about the data format, the release version of the software with which they were created and the corresponding input files. Because individual TMBs are too small to efficiently access them from the tape system around 10 of them need to be merged into a single file. This complicates the work-flow.

In total 300M events or 75TB of raw data needed to be processed to create an expected 45TB of reconstructed output data (DSTs and TMBs). Given a reconstruction speed of around 50s/Event the estimated need for CPU time was 4 million hours or nearly 500 years on a 1GHz PIII machine. With a desired completion time of 3 months nearly 2000 CPUs were required.

The DØ reconstruction farm (ca. 1000 such CPUs) was available only during shutdown times as it is dimensioned only to keep pace with data-taking. Thus other resources had to be used to achieve a timely completion of the task. The resources available to DØ through paritcipating institutes are listed in Tab. 1. Remote clusters, however, have a reduced network bandwidth and an increased IP round trip time to the central systems serving the databases and input files.

In first tests the reduced network connectivity showed catastrophic performance problems. Processing individual events would sometimes take up to an hour instead of minutes. The problem arose from a huge number of accesses to the calibration database repeated for each event. This influenced the operational mode strongly.

## 3. Operation

Following the requirements imposed on the software by the available resources the plan for operation was modified not to use database access. In this mode the DSTs of a previous processing are used as input which reduces the amount of data to be read by 50%. Due to limited person-power the merging of TMB results was done centrally, where corresponding scripts exsisted with sufficient cross-checking to avoid duplication.

For data distribution SAM[5] was used. Datasets were created manually, assigned to the sites and published on a web page. An operator at each site would start the actual jobs for the assigned datasets. Most sites pre-staged these datasets into the sites SAM cache which was reported to be less error prone.

Once the jobs finished the output TMBs, the corresponding metadata and a defined set of log-files were copied to a central location. The transport was based on the `sam_cp` interface provided by SAM and a rudimentary file based book-keeping that avoids repeated transport and indicates successful completion to the central site was applied. The TMBs were then merged and stored into SAM.

Beside the manual job submission at individual sites the three British sites were operated from a single entry point using the Globus[6] and Condor[7] based tool JIM[8,9,10] to distribute the jobs among the three sites.

Before a site was allowed to join the official production running it had to be certified. Each site processed one well defined dataset of 66 files. A large number of plots produced from the results were compared between the sites as well as to a production done from raw.

## 4. Conclusions and Outlook

DØ performed the first HEP-data re-reconstruction on globally distributed, non-dedicated clusters. 300M events or 45TB of input data were processed at seven centres. Around 30% or 15TB were processed outside of Fermilab using Grid concepts for data- and job-distribution as they are foreseen for the LHC Computing Grid (LCG).

The successful completion of the DØ reprocessing has proven these concepts to be functional, but also revealed problems that need to be addressed for a fully automated production. The most important of these being the automatic detection and recovery from job failures which don't have their cause in the program itself, but arise from a temporary failure in the environment.

## References

1. *Daniel Wicke and Mike Diesburg.* `http://www-d0.fnal.gov/computing/reprocessing/`, 2004.
2. *K. Riesselmann. Ferminews* **27, 2**(2004) 2. `http://www.fnal.gov/pub/ferminews/ferminews04-02-01/p1.html`.
3. *M. Diesburg and D. Wicke.* Contribution #441 to CHEP, 27th Sep. to 1st Oct. 2004, Interlaaken, Switzerland, 2004.
4. *K. Riesselmann. CERN Courier* **44, 7**(2004) 16. `http://www.cerncourier.com/main/article/44/7/15`.
5. DØ-Coll. Collaboration, *Mike Diesburg et al.* DØ-Note 3464, 1998.
6. *I. Foster and C. Kesselman. Intl J. Supercomputer Applications* **11(2)**(1997) 115–128.
7. *D.H.J. Epema, M. Livny, R. van Dantzig, X. Evers, and J. Pruyne. Future Generation Computer Systems* **12**(1996) 53–65.
8. *A. Baranovski et al. Nucl. Phys. Proc. Suppl.* **120**(2003) 119–125.
9. *A. Baranovski et al. Nucl. Instrum. Meth.* **A502**(2003) 423–425.
10. *A. Baranovski et al.* (2003) . ArXiv:cs.dc/0307007.